HIERARCHICAL DOMAIN GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study hierarchical domain generalization as a problem of extrapolation from finite observed regions to an entire instance space, replacing i.i.d. sampling with arbitrary domain hierarchies. We show that the central obstruction is not only the complexity of the hypothesis class, but the train/test domain partition through which evidence is revealed. In particular, on an infinite domain, for any class with more than two hypotheses and any training size, some partition makes generalization fail for some target. These results suggest that modern generalization theory must treat domain structure as a first-class object.

1 INTRODUCTION

The concept of generalization in machine learning has been shaped by the seminal work of Vapnik and Chervonenkis on statistical generalization theory (Vapnik & Chervonenkis, 1971) and by Valiant's introduction of the PAC model for binary classification (Valiant, 1984).

In these setups, training data is drawn i.i.d. from some unknown data distribution $\mathcal{D}$, and the learner's output is evaluated on the *same* $\mathcal{D}$. A successful learner is required to achieve low expected error on the distribution that generated the training data (or related data distributions in more recent work on domain adaptation learning, e.g., Ben-David et al. (2006)). This statistical framework has shaped the field's conception of "what it means to generalize" for four decades.

Motivated by length generalization in language models, we study exact extrapolation from finite observed regions to an entire input space. For example, a model trained on five-digit addition may fit its training data perfectly yet fail on six-digit numbers, as it often does in practice (Anil et al., 2022; Lee et al., 2024). This motivates asking when observations on a restricted input region suffice for eventual recovery of the underlying rule.

We study this question in a realizable, noiseless model of exact identification along a domain hierarchy. Instead of measuring expected error under a data-generating distribution, we require the learner to recover the target function on the entire instance space Ω. Instead of assuming an i.i.d. observation model, we let the observable region grow through an externally specified hierarchy $\bar{\mathcal{X}} = (\mathcal{X}_n)_{n\geq 0}$ of finite cumulative domains, with $\emptyset = \mathcal{X}_0 \subsetneq \mathcal{X}_1 \subsetneq \mathcal{X}_2 \subsetneq \cdots$ and $\bigcup_{n\geq 1} \mathcal{X}_n = \Omega$ (so Ω is countable). At level n, the learner observes the labeled training set $\{(x, f(x)) : x \in \mathcal{X}_n\}$ and is evaluated on all of Ω; equivalently, it must extrapolate from the observed region $\mathcal{X}_n$ to the unobserved complement $\Omega \setminus \mathcal{X}_n$. The only way to achieve this is to exactly *identify* f.

This captures a common desideratum of modern reasoning systems: train on shorter inputs and generalize to longer ones (Anil et al., 2022; Lee et al., 2024; Zhou et al., 2024) (a.k.a. length generalization), train on simpler instances and generalize to harder ones (Bengio et al., 2009), and train on bounded-horizon tasks and generalize to tasks with longer horizons (Abbe et al., 2023; Yang et al., 2026b). In the point-by-point case, where each increment is a singleton $\mathcal{X}_n \setminus \mathcal{X}_{n-1} = \{x_n\}$ and hence $\mathcal{X}_n = \{x_1, \ldots, x_n\}$, the labels form a nonrepeating informant presentation with both positive and negative examples in Gold's sense (Gold, 1967).

The central question is: given a hierarchy $\bar{\mathcal{X}}$, how many levels N must a learner observe to recover the target $f \in \mathcal{H}$ on the entire domain Ω? We study this question under different uniformity requirements and relate the resulting guarantees to classical identification, online learning, and length generalization.

We characterize four identification criteria according to which objects the required level may depend on. If the level may depend on both the target and the hierarchy, countability of $\mathcal{H}$ is exactly the

right condition. If the hierarchy is fixed but the same level must work for every target, the threshold becomes finiteness of $\mathcal{H}$. But if the level must be chosen uniformly over the hierarchy itself, the situation changes completely: for every binary hypothesis class with more than two hypotheses, such a guarantee is possible only in the finite-domain case (Theorems 1, 3 and 6).

Classical identification in the limit asks whether the learner eventually succeeds; quantitative length generalization asks how far along the length hierarchy it must observe. The hierarchy is string length: training on all strings up to length N and testing on longer strings asks whether that observed prefix determines the target rule. Since relevant classes such as finitely encoded Transformer families are infinite, we stratify them by complexity: for each bound s, we ask how many lengths are needed to identify every target of complexity at most s. Existing non-asymptotic length-generalization results estimate this sublevel domain complexity for particular representation classes and complexity scales (Chen et al., 2025; Yang et al., 2026a); qualitative and approximate variants fit the same fixed-hierarchy picture (Huang et al., 2025; Izzo et al., 2026).

The dependence of these bounds on the hierarchy is unavoidable. Even with the hypothesis class and complexity scale fixed, another hierarchy can make the finite-slice bounds grow arbitrarily fast (Theorem 10 and Corollary 11). Thus a length-generalization bound is a bound for the length hierarchy, not a hierarchy-free guarantee for arbitrary domain shifts.

Together, these results distinguish three kinds of claims: qualitative guarantees that follow from cardinality, hierarchy-uniform guarantees ruled out by impossibility results, and quantitative bounds that require analysis of a specific class and hierarchy. Classical no-free-lunch theorems warn that an unrestricted hypothesis class admits no uniform learning guarantee (Shalev-Shwartz & Ben-David, 2014). Here, even after restricting the class, the order in which distinguishing evidence becomes available remains essential.

A uniform online mistake bound need not imply target-uniform identification: a class can admit at most one prediction mistake yet have no common identification level across targets (Example 1). Under positive-only presentations, bounded pairwise overlap permits uniform identification even for infinite classes. We develop these comparisons and their relation to generation in the limit (Kleinberg & Mullainathan, 2024) in Section E. Proposition 16 shows that hierarchy-uniform identification can remain impossible at a positive error tolerance under a fixed full-support evaluation distribution. Distributional support mismatch is treated in Section E.3; noisy labels and model misspecification remain outside our scope.

Organization. Sections 2 and 3 give the framework and characterizations. Sections 4 and 5 develop quantitative bounds and the length-generalization connection. Section 6 establishes the hierarchy dependence. Related learning settings are treated in the appendix (Section E).

2 Hierarchical Domain Generalization

2.1 Setup

Let Ω be a countably infinite instance space and let $\mathcal{H} \subseteq \{0,1\}^\Omega$ be a binary hypothesis class.[1] We work in the realizable setting with deterministic labeling: an unknown target $f \in \mathcal{H}$ generates labels $f(x) \in \{0,1\}$ for $x \in \Omega$. For any subset $\mathcal{X} \subseteq \Omega$, write $f|_{\mathcal{X}}$ for the labeled training set $\{(x, f(x)) : x \in \mathcal{X}\}$.

To model out-of-domain generalization, we must specify how the observable part of Ω grows. The primitive object is not a sampling distribution, but a predetermined sequence of finite cumulative regions whose labels become available to the learner.

Definition 1 (Domain hierarchy)**.** *A* hierarchy *on Ω is a telescopic exhaustion $\bar{\mathcal{X}} = (\mathcal{X}_n)_{n\geq 0}$ by finite cumulative observed regions satisfying*

$$\emptyset = \mathcal{X}_0 \subsetneq \mathcal{X}_1 \subsetneq \mathcal{X}_2 \subsetneq \cdots, \qquad |\mathcal{X}_n| < \infty \text{ for every } n, \qquad \bigcup_{n\geq 1} \mathcal{X}_n = \Omega. \tag{1}$$

For $n \geq 1$, write $\Delta_n := \mathcal{X}_n \setminus \mathcal{X}_{n-1}$ for the newly revealed increment at level n. The definition is intentionally agnostic about how the regions are ordered. One common source is a sequence of

[1] Finite domains are used only as degenerate comparison cases, with the convention that the hierarchy collapses to $\mathcal{X}_n = \Omega$ for all large enough n. We state the characterizations for binary labels.

externally specified domains, such as different populations, demographic subgroups, or intervention-based environments. Another source orders Ω by an intrinsic property of the inputs, such as length or difficulty, placing simpler inputs in early levels and harder ones later, as in length generalization and curriculum learning. When each increment is a singleton, say $\Delta_n = \{x_n\}$, equivalently $\mathcal{X}_n = \{x_1, \ldots, x_n\}$, the labels form a nonrepeating informant presentation in Gold's sense (Gold, 1967).

Hierarchical Domain Generalization. Once a hierarchy is fixed, level n splits Ω into the observed region $\mathcal{X}_n$, from which the learner receives the labeled training set $f|_{\mathcal{X}_n}$, and the unobserved complement $\Omega \setminus \mathcal{X}_n$, which is out of domain relative to that level. The learner is evaluated on the entire Ω, so success requires extrapolation from the observed region to the unseen part of the hierarchy. This in-domain/out-of-domain split is determined by $\bar{\mathcal{X}}$ alone, without any distributional assumptions.

We idealize each observed level as fully labeled: at level n, the learner has access to all labels on $\mathcal{X}_n$. We ask whether the learner eventually stabilizes on the correct target, without requiring the observed labels to uniquely determine it.

Definition 2 (Learner)**.** *A learning algorithm is a map $A : \bigcup_{\mathcal{X} \subseteq \Omega,\ |\mathcal{X}| < \infty} \{0,1\}^{\mathcal{X}} \to \{0,1\}^{\Omega}$, which takes the labeled training set $f|_{\mathcal{X}}$ as input and returns a binary predictor $A(f|_{\mathcal{X}})$. The learner may be chosen with knowledge of the hypothesis class $\mathcal{H}$.*

The learner sees the labeled set, not its presentation order, and need not be computable.

Definition 3 (Domain complexity)**.** *Given a learner A, a target $f \in \mathcal{H}$, and a hierarchy $\bar{\mathcal{X}}$, the* domain complexity *of A on f along $\bar{\mathcal{X}}$ is*

$$N_A^{\bar{\mathcal{X}}}(f) := \min \left\{ i \in \mathbb{N}_{\geq 1} : \forall n \geq i,\ A(f|_{\mathcal{X}_n}) = f \right\}, \tag{2}$$

with value ∞ if no such finite i exists.

Thus $N_A^{\bar{\mathcal{X}}}(f)$ is the first level after which, for every $n \geq N_A^{\bar{\mathcal{X}}}(f)$, training on the observed region $\mathcal{X}_n$ yields a hypothesis correct on all of Ω, equivalently on every later and currently unobserved part of the hierarchy.

2.2 Levels of Learning Goals

We distinguish the possible forms of hierarchical domain generalization by the order of the quantifiers over the learner, the target, the hierarchy, and the identifying level. The first two criteria are hierarchy-dependent: in **A1**, the identifying level may depend on both the target and the hierarchy; in **A2**, it may still depend on the hierarchy, but must be common to all targets in the class.

Definition 4 (**A1**: *Non-uniform identifiability*)**.** *$\mathcal{H}$ is* non-uniformly identifiable *(**A1**) if there exists a learning algorithm A with the following property: for every target $f \in \mathcal{H}$ and every hierarchy $\bar{\mathcal{X}}$, there exists a finite N such that $N_A^{\bar{\mathcal{X}}}(f) \leq N$.*

A2 keeps the dependence on the hierarchy but requires one level to work for all targets.

Definition 5 (**A2**: *Target-uniform identifiability*)**.** *$\mathcal{H}$ is* target-uniformly identifiable *(**A2**) if there exists a learning algorithm A with the following property: for every hierarchy $\bar{\mathcal{X}}$, there is a finite N such that, for every target $f \in \mathcal{H}$, $N_A^{\bar{\mathcal{X}}}(f) \leq N$.*

The remaining criteria are domain-uniform. Unlike **A1** and **A2**, they require the level bound to be chosen without knowing the hierarchy, and hence without knowing the induced train/test split. This is the hierarchy analogue of choosing a PAC sample-size bound before the data-generating distribution is revealed.

Definition 6 (**A3**: *Domain-uniform identifiability*)**.** *$\mathcal{H}$ is* domain-uniformly identifiable *(**A3**) if there exists a learning algorithm A with the following property: for every target $f \in \mathcal{H}$, there exists a finite N such that, for every hierarchy $\bar{\mathcal{X}}$, $N_A^{\bar{\mathcal{X}}}(f) \leq N$.*

A3 still permits target-dependent bounds. **A4** removes this last dependence: it requires a single finite level that works for all targets and all hierarchies. In this sense, **A4** is the fully uniform analogue of a sample-complexity guarantee.

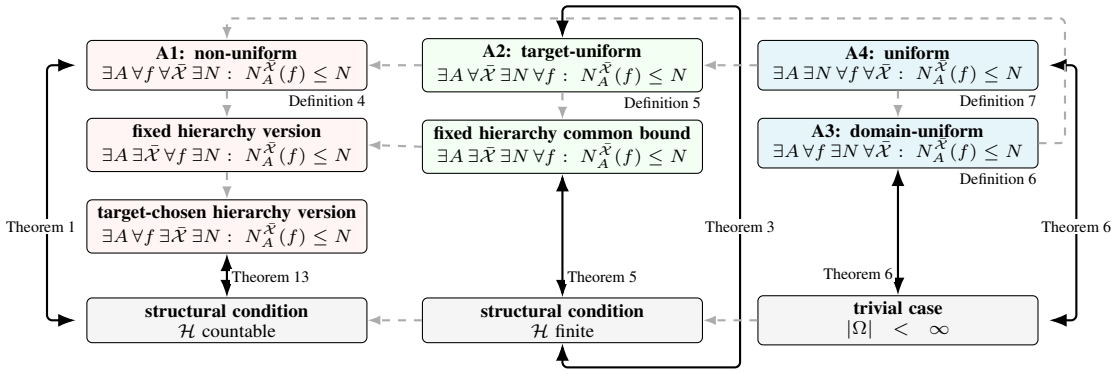

Figure 1: **Identifiability landscape.** The four learning criteria and their equivalent structural conditions, for $f \in \mathcal{H}$, arbitrary hierarchies $\bar{\mathcal{X}}$, and $|\mathcal{H}| > 2$. Gray dashed arrows denote formal implications; black double arrows denote substantive equivalences. Transitive arrows are omitted.

Definition 7 (**A4**: *Uniform identifiability*)**.** *$\mathcal{H}$ is* uniformly identifiable *(**A4**) if there exist a learning algorithm A and a finite N with the following property: for every target $f \in \mathcal{H}$ and every hierarchy $\bar{\mathcal{X}}$, $N_A^{\bar{\mathcal{X}}}(f) \le N$.*

3 Exact Characterization of Domain Generalization

We now characterize the four learning goals from Section 2.2.

3.1 Non-Uniform Domain-Dependent Generalization

Very little, it turns out, is required for the weakest criterion. If we settle for **A1**, $\mathcal{H}$ being countable is all we need.

Theorem 1 (**A1** $\Leftrightarrow$ Countable $\mathcal{H}$)**.** *$\mathcal{H}$ is non-uniformly identifiable (**A1**, Definition 4) if and only if $\mathcal{H}$ is countable.*

Proof sketch. Enumerate $\mathcal{H}$ and let A be the consistent learner with this fixed tie-breaking order: after observing a prefix, it outputs the first hypothesis in the enumeration that is consistent with the observed labels. For $f = h_m$, each lower-index h_j is eventually ruled out as $\bar{\mathcal{X}}$ reveals its disagreement with f, leaving h_m as the first such hypothesis. Conversely, fix any hierarchy. If a learner identifies every target after some finite prefix, then each target is associated with at least one finite labeled prefix on which the learner outputs it. Since there are only countably many finite labeled prefixes, $\mathcal{H}$ must be countable. □

The countability characterization is still learner-dependent: it says that some chosen rule eventually outputs the target, not that the observed labels have ruled out every other hypothesis. A learner may identify the target by choosing one hypothesis consistent with the observed labels while other hypotheses still agree with those labels. The following example makes this separation explicit.

Proposition 2. *There exist a countable hypothesis class $\mathcal{H}$, a hierarchy $\bar{\mathcal{X}}$, and a successful learner A for $\mathcal{H}$ over $\bar{\mathcal{X}}$ (namely, for every $f \in \mathcal{H}$, $N_A^{\bar{\mathcal{X}}}(f) < \infty$) so that for some $f^* \in \mathcal{H}$ for every finite level n, there exists some $g_n \in \mathcal{H}$ so that $g_n|_{\mathcal{X}_n} = f^*|_{\mathcal{X}_n}$, and yet $g_n \neq f^*$ (namely, for some $x \in \Omega$, $f^*(x) \neq g_n(x)$).*

Proof sketch. Let $\Omega = \mathbb{N}$ and $\mathcal{H}$ contain the all-zero function 0 together with the singleton indicators $h_i = \mathbb{1}_{\{i\}}$. Take the natural hierarchy $\mathcal{X}_n = \{1, \dots, n\}$ and the target $f = 0$. A least-index learner that lists 0 first identifies 0 immediately, so $N_A^{\bar{\mathcal{X}}}(0) < \infty$. But for every n and every $j > n$, the singleton h_j agrees with 0 on $\mathcal{X}_n$. Thus no finite prefix by itself rules out all alternatives to 0. □

Finite isolation, which characterizes when the observations uniquely determine the target, is hierarchy-invariant; see Proposition 12 in the appendix.

The same countability threshold holds when the hierarchy is fixed in advance or chosen separately for each target; Theorem 13 in the appendix gives the equivalent formulations.

3.2 Target-Uniform Domain-Dependent Generalization

A2 asks for a different kind of success. Once the hierarchy is fixed, the same prefix level must work for every target: the amount of evidence is independent of which $f \in \mathcal{H}$ generated the labels.

Theorem 3 (**A2** $\Leftrightarrow$ Finite $\mathcal{H}$)**.** *$\mathcal{H}$ is target-uniformly identifiable (**A2**, Definition 5) if and only if $\mathcal{H}$ is finite.*

Proof sketch. With $\mathcal{H}$ finite, the least-index learner from Theorem 1 achieves a common bound on any hierarchy: $\max_{f\in\mathcal{H}} N_A^{\bar{\mathcal{X}}}(f) < \infty$. Conversely, a common bound N on a single hierarchy makes the labeled-prefix map $f \mapsto f|_{\mathcal{X}_N}$ injective on $\mathcal{H}$; otherwise two targets with the same prefix would give the learner the same input at level N. Hence $|\mathcal{H}| \le 2^{|\mathcal{X}_N|}$. □

The proof isolates the mechanism behind **A2**: at a common level, two targets with the same prefix are indistinguishable to any learner. The relevant quantity is therefore the first prefix that separates the class.

Definition 8 (Class domain complexity)**.** *For a hypothesis class $\mathcal{H}$ and a hierarchy $\bar{\mathcal{X}}$, the class domain complexity is*

$$N^{\bar{\mathcal{X}}}(\mathcal{H}) := \min\{n \in \mathbb{N}_{\ge 1} : \forall f, g \in \mathcal{H},\ f \neq g \Longrightarrow \exists x \in \mathcal{X}_n : f(x) \neq g(x)\}. \tag{3}$$

If no such finite level exists, set $N^{\bar{\mathcal{X}}}(\mathcal{H}) := \infty$.

Proposition 4 (Optimality of class domain complexity)**.** *For every hierarchy $\bar{\mathcal{X}}$, $N^{\bar{\mathcal{X}}}(\mathcal{H}) = \min_A \sup_{f\in\mathcal{H}} N_A^{\bar{\mathcal{X}}}(f)$, where A ranges over learners and both sides take values in $\mathbb{N}_{\ge 1} \cup \{\infty\}$.*

Thus the optimal fixed-hierarchy **A2** bound is algorithm-agnostic, and the variants below are equivalent: a common bound along one hierarchy already forces $\mathcal{H}$ to be finite, while finiteness gives a common bound along every hierarchy.

Theorem 5 (Equivalent conditions for **A2**)**.** *The following are equivalent:*

(i) $\exists A\ \forall \bar{\mathcal{X}}\ \exists N \in \mathbb{N}_{\ge 1}\ \forall f \in \mathcal{H} :\ N_A^{\bar{\mathcal{X}}}(f) \le N$ *(that is, $\mathcal{H}$ satisfies **A2**, Definition 5);*
(ii) $\exists A\ \exists \bar{\mathcal{X}}\ \exists N \in \mathbb{N}_{\ge 1}\ \forall f \in \mathcal{H} :\ N_A^{\bar{\mathcal{X}}}(f) \le N$*;*
(iii) $\exists \bar{\mathcal{X}} :\ N^{\bar{\mathcal{X}}}(\mathcal{H}) < \infty$*;* (iv) $\forall \bar{\mathcal{X}} :\ N^{\bar{\mathcal{X}}}(\mathcal{H}) < \infty$*;* (v) *$\mathcal{H}$ is finite.*

3.3 Domain-Uniform Generalization

Theorem 6 (**A3**/**A4** characterization)**.** *Assume $|\mathcal{H}| > 2$. The following are equivalent:*

*(i) $\mathcal{H}$ is uniformly identifiable (**A4**, Definition 7);*
*(ii) $\mathcal{H}$ is domain-uniformly identifiable (**A3**, Definition 6);*
(iii) Ω is finite.

Proof sketch. (i) implies (ii). This is immediate. *(iii) implies (i).* If Ω is finite, then every hierarchy reveals all of Ω by level $|\Omega|$. At that level the observed labels determine the target on the entire domain. A learner that outputs the unique hypothesis in $\mathcal{H}$ consistent with the full-domain labels identifies every target under every hierarchy by level $|\Omega|$.

(ii) implies (iii). Suppose **A3** holds through a learner A and target-dependent levels $(N_f)_{f\in\mathcal{H}}$. If two distinct hypotheses f, g agreed on at least $M = \max\{N_f, N_g\}$ points, choose a finite set $S \subseteq \{f = g\}$ of size M and a hierarchy with $\mathcal{X}_M = S$. Since $f|_S = g|_S$, the learner receives the same input under both targets at level M, but **A3** requires the same output to equal both f and g, a contradiction. Hence every pairwise agreement set is finite.

Since $|\mathcal{H}| > 2$, choose three distinct hypotheses $h_1, h_2, h_3 \in \mathcal{H}$. Write $\{h_i = h_j\} := \{x \in \Omega : h_i(x) = h_j(x)\}$. At each $x \in \Omega$, two of the three binary labels must agree, so

$$\{h_1 = h_2\} \cup \{h_1 = h_3\} \cup \{h_2 = h_3\} = \Omega. \tag{4}$$

The three sets on the left are finite by the previous paragraph. Hence Ω is finite. □

Under the standing assumption of this paper, Ω is countably infinite. Thus for every class with $|\mathcal{H}| > 2$, both **A3** and **A4** are impossible. The only remaining cases are degenerate. Singleton classes are trivial. For a two-element class $\mathcal{H} = \{h, g\}$, domain-uniform (**A3**), equivalently uniform (**A4**), identification is possible exactly when h and g agree on only finitely many points: then any sufficiently large observed region contains a point where they differ, and the label at that point identifies the target.

Why Domain Generalization Is Hard. The obstruction above is not a usual PAC-learning obstruction from VC dimension, class size, or agnostic noise. It comes from the train/test partition itself. The useful way to read the obstruction is adversarial:

Suppose Ω is infinite and take any three distinct hypotheses $h_1, h_2, h_3 \in \mathcal{H}$. Then some pair must agree on infinitely many points. Indeed, if both $\{h_1 = h_2\}$ and $\{h_1 = h_3\}$ are finite, then their union is finite, so on infinitely many points h_2 and h_3 both take the binary label opposite to h_1. Hence $\{h_2 = h_3\}$ is infinite. Thus, from any class with $|\mathcal{H}| > 2$, an adversary can choose two distinct hypotheses f, g with an infinite agreement set and place arbitrarily many of those agreement points at the beginning of the hierarchy. On any such prefix $S \subseteq \{f = g\}$, the learner receives the same labeled training set under the two possible targets. Whatever it outputs on that input, at least one of f and g is not identified.

4 Stratified Domain Complexity

The characterizations above are qualitative: finite classes admit target-uniform bounds along a fixed hierarchy, while infinite classes do not. To study infinite model families quantitatively, we stratify them by model complexity and ask how the best bound grows with the allowed complexity.

Complexity Sublevels. A natural refinement is to keep track of model complexity. Many representation classes of interest are not just infinite sets of functions: programs, automata, grammars, and implemented neural-network families come with finite descriptions or size parameters, such as description length,[2] program length, number of states, grammar size, depth, width, or number of parameters. We require the chosen scale to have finite sublevels; bounded description length suffices, whereas neural-network depth, width, or parameter count alone need not.

We now formalize this fixed-hierarchy quantity. Fix a hierarchy $\bar{\mathcal{X}}$ on Ω, a hypothesis class $\mathcal{H} \subseteq \{0,1\}^\Omega$, and a model-complexity function $c : \mathcal{H} \to \mathbb{N}$. For $s \in \mathbb{N}$, the *sublevel class* of model complexity at most s is

$$\mathcal{H}^c_{\le s} := \{f \in \mathcal{H} : c(f) \le s\}. \tag{5}$$

Throughout this section, we consider complexity scales whose sublevels are finite and exhaustive:

$$|\mathcal{H}^c_{\le s}| < \infty \text{ for every } s \in \mathbb{N}, \qquad \bigcup_{s \in \mathbb{N}} \mathcal{H}^c_{\le s} = \mathcal{H}. \tag{6}$$

Sublevel Bounds. Abstractly, every countable class admits such a stratification by enumeration. In applications, however, the scale c is chosen because it reflects a representation of the hypotheses. The quantity of interest is the best target-uniform bound on each sublevel class.

Definition 9 (Sublevel domain complexity)**.** *The* sublevel domain complexity *of* $(\mathcal{H}, c)$ *along* $\bar{\mathcal{X}}$ *is the function* $N^{\bar{\mathcal{X}}}_{(\mathcal{H},c)} : \mathbb{N} \to \mathbb{N}_{\ge 1} \cup \{\infty\}$ *defined by*

$$N^{\bar{\mathcal{X}}}_{(\mathcal{H},c)}(s) := N^{\bar{\mathcal{X}}}(\mathcal{H}^c_{\le s}) = \min\left\{i \in \mathbb{N}_{\ge 1} : \forall h \ne h' \in \mathcal{H}^c_{\le s},\ \exists x \in \mathcal{X}_i \text{ s.t. } h(x) \ne h'(x)\right\}. \tag{7}$$

If no such i *exists, set* $N^{\bar{\mathcal{X}}}_{(\mathcal{H},c)}(s) := \infty$.

[2] For a prefix-code description length, Kraft's inequality implies that only finitely many descriptions have code length at most any fixed s. We use only this finite-slice consequence, not an MDL or probabilistic argument.

Under the finite-sublevel assumption, each sublevel class has some finite target-uniform identifying level along the fixed hierarchy. The next result shows that the smallest possible such bound is exactly the sublevel domain complexity just defined.

Theorem 7 (Sublevel domain complexity is the optimal rate)**.** *Fix a hierarchy $\bar{\mathcal{X}}$, a class $\mathcal{H} \subseteq \{0,1\}^{\Omega}$, and a complexity scale $c : \mathcal{H} \to \mathbb{N}$ satisfying Equation* (6)*. Then, for every $s \in \mathbb{N}$ with $\mathcal{H}^c_{\leq s} \neq \emptyset$,*

$$N^{\bar{\mathcal{X}}}_{(\mathcal{H},c)}(s) = \min_{A} \sup_{f \in \mathcal{H}^c_{\leq s}} N^{\bar{\mathcal{X}}}_{A}(f), \tag{8}$$

where A ranges over learners. If, in addition, $|\mathcal{H}^c_{\leq s}| \to \infty$ as $s \to \infty$, then $N^{\bar{\mathcal{X}}}_{(\mathcal{H},c)}(s) \to \infty$.

Thus finite sublevels give the qualitative guarantee, while Theorem 7 identifies the best possible target-uniform bound at each complexity level. When the sublevel classes grow, these bounds must also diverge. The remaining interesting question is how fast the sublevel domain complexity function grows.

5 Length Generalization as a Case Study

Length generalization is the fixed-hierarchy case in which the hierarchy is ordered by input length and the quantitative question is how sublevel domain complexity grows with model complexity.

Concretely, length generalization asks whether a model trained on short inputs succeeds on longer inputs. This question is especially visible for Transformers (Vaswani et al., 2017) trained on algorithmic tasks, where models can fit all short instances yet fail sharply on longer ones (Anil et al., 2022; Lee et al., 2024; Zhou et al., 2024).

Definition 10 (Length hierarchy)**.** *Formally, fix a finite alphabet Σ and consider the hierarchy*

$$\Omega := \Sigma^+, \qquad \bar{\Sigma} := (\Sigma^{\leq N})_{N \geq 0}, \qquad \Sigma^{\leq 0} := \emptyset, \qquad \Sigma^{\leq N} := \Sigma \cup \Sigma^2 \cup \cdots \cup \Sigma^N \quad (N \geq 1). \tag{9}$$

A learner trained on lengths at most N observes all labels on $\Sigma^{\leq N}$ and is evaluated on all of Σ^+. Thus length generalization is exact identification under the fixed cumulative hierarchy $\bar{\Sigma}$.

We write the representation class as $\mathcal{H} \subseteq \{0,1\}^{\Sigma^+}$.[3] Here $\mathcal{H}$ may be a class of languages, programs, or Transformer-realizable predictors, and c may encode description length, program size, Transformer depth, number of heads, parameter precision, norm bounds, or related architectural parameters. The quantitative content of non-asymptotic length generalization lies in estimating the growth of $N^{\bar{\Sigma}}_{(\mathcal{H},c)}(s)$ for concrete choices of $(\mathcal{H}, c)$.

Existing Results. Huang et al. (2025) give a qualitative fixed-hierarchy, target-wise guarantee for Transformer length generalization: under an idealized regularized inference procedure, targets in their Limit-Transformer class are eventually identified along $\bar{\Sigma}$. This does not estimate a sublevel domain-complexity growth rate.

More quantitatively, Chen et al. (2025) estimate this sublevel domain complexity for several representation classes. For DFAs, the model class is regular languages and the complexity measure is the number of states; their bound is $2q - 2$ for targets recognized by DFAs with at most q states, including the empty string among the observed inputs.[4] For linear CFGs, the complexity measure is grammar description size, and the analogous sublevel bound has no computable upper bound in that size. For restricted C-RASP,[5] the relevant parameters are precision T in the one-layer Transformer setting, and precision T together with K heads in the two-layer setting; the resulting bounds are $O(T^2)$ and $O(T^{O(K)})$, respectively.

[3] The binary-valued language notation is only for convenience; the same setup extends directly to finite-output maps such as functions $\Sigma^* \to \Sigma$.

[4] For our nonempty-string hierarchy, $2q - 1$ suffices: apply the bound to the automata reached after the first symbol.

[5] C-RASP programs compile to softmax Transformers (Yang & Chiang, 2024); hence C-RASP gives a programmable lower bound on softmax-Transformer expressivity and is often used as a proxy for studying Transformer expressiveness.

Later work continues the same calculation for more Transformer-facing classes. Yang et al. (2026a) study the same sublevel domain-complexity function for C-RASP, proving uncomputability for full C-RASP and tight exponential growth for positive C-RASP. Izzo et al. (2026) move to an approximate analogue: exact identification is replaced by approximate agreement on all longer inputs, with bounds in terms of norms, locality, periodicity, precision, and margins.

Beyond Length. The preceding results all concern the fixed length hierarchy $\bar{\Sigma}$. The uncomputability statements in this literature should not be interpreted as a peculiarity of ordering strings by length. The obstruction can already come from the representation of $\mathcal{H}$: if semantic equivalence of descriptions is undecidable, then the sublevel domain-complexity function has no computable upper bound in the complexity parameter s.

To discuss computability, we use a represented class: a finite description $p \in \Pi$ denotes a computable hypothesis $h_p : \Omega \to \{0,1\}$, and $\kappa : \Pi \to \mathbb{N}$ is a computable complexity scale on descriptions. A single algorithm must compute $h_p(x)$ from (p, x) for every $p \in \Pi$ and $x \in \Omega$. For a complexity bound s and hierarchy $\bar{\mathcal{X}}$, write

$$\mathcal{H}^{\Pi,\kappa}_{\leq s} := \{h_p : p \in \Pi,\ \kappa(p) \leq s\}, \qquad N^{\bar{\mathcal{X}}}_{(\Pi,\kappa)}(s) := N^{\bar{\mathcal{X}}}(\mathcal{H}^{\Pi,\kappa}_{\leq s}). \tag{10}$$

The quantity on the right is still the information-theoretic domain complexity from Definition 9. A hierarchy is computable if its finite levels can be listed, and two descriptions are semantically equivalent if their denoted hypotheses agree on all of Ω.

Proposition 8 (Computability obstruction)**.** *If semantic equivalence is undecidable for descriptions in Π, then for every computable exhaustive hierarchy $\bar{\mathcal{X}}$ on Ω, there is no computable function $B : \mathbb{N} \to \mathbb{N}$ such that*

$$N^{\bar{\mathcal{X}}}_{(\Pi,\kappa)}(s) \leq B(s) \qquad \textit{for all } s \in \mathbb{N}, \tag{11}$$

For classes such as linear CFGs, this recovers the absence of a computable non-asymptotic bound along $\bar{\Sigma}$ shown by Chen et al. (2025), and holds for any computable exhaustive hierarchy. The computability requirement on the hierarchy is essential: without it, the hierarchy itself can encode the finite witnesses that separate the relevant hypotheses.

Proposition 9 (Necessity of computability)**.** *Let Ω be countably infinite, and let $\mathcal{H} = \{h_1, h_2, \ldots\} \subseteq \{0,1\}^\Omega$ be countable. There exists an exhaustive hierarchy $\bar{\mathcal{X}}$ such that $N^{\bar{\mathcal{X}}}(\{h_1, \ldots, h_m\}) \leq m$ for all $m \geq 1$.*

Thus, for a complexity scale c with finite sublevels, if the enumeration is ordered so that $\mathcal{H}^c_{\leq s} \subseteq \{h_1, \ldots, h_{b(s)}\}$ for some computable b, the same hierarchy gives $N^{\bar{\mathcal{X}}}_{(\mathcal{H},c)}(s) \leq b(s)$ for all s. Hence the computability assumption in Proposition 8 is essential.

Together, these observations separate two sources of difficulty. The computability obstruction above shows that non-computable rates can arise from the effective description of $\mathcal{H}$ itself, provided the hierarchy is computable. The next section concerns the other source: even with $(\mathcal{H}, c)$ fixed, the growth of the sublevel domain complexity need not be determined by this pair alone.

6 No Free Lunch for Domain Generalization

The preceding sections give meaningful quantitative notions once a hierarchy is fixed. For a target, a learner has an identification level $N^{\bar{\mathcal{X}}}_A(f)$; for a finite class, $N^{\bar{\mathcal{X}}}(\mathcal{H})$ is the optimal target-uniform bound; and for a complexity scale c, $N^{\bar{\mathcal{X}}}_{(\mathcal{H},c)}(s)$ is the corresponding optimal bound on the sublevel class $\mathcal{H}^c_{\leq s}$. All three are defined relative to the hierarchy $\bar{\mathcal{X}}$.

This dependence is essential. Theorem 6 already shows that, for classes with $|\mathcal{H}| > 2$, domain generalization uniformly over all hierarchies is possible only in the finite-domain case. We now formalize the same dependence as a no-free-lunch result: over an infinite domain, changing the hierarchy alone can make each of the fixed-hierarchy quantities above arbitrarily large.

Class-Level Bounds. We first ask this question before adding a complexity scale.

Theorem 10 (No free lunch for domain generalization)**.** *Let Ω be countably infinite. For any $\mathcal{H} \subseteq \{0,1\}^{\Omega}$ with $|\mathcal{H}| > 2$ and every $M \in \mathbb{N}_{\geq 1}$, there exists a hierarchy $\bar{\mathcal{X}}$, even one revealing a single new point at each level, such that, for every learner A, there exists a target $h \in \mathcal{H}$ with*

$$N_A^{\bar{\mathcal{X}}}(h) > M. \tag{12}$$

When $\mathcal{H}$ is finite, this is equivalent to the class-domain-complexity statement that, for every $M \in \mathbb{N}_{\geq 1}$, there exists a hierarchy $\bar{\mathcal{X}}$ such that $N^{\bar{\mathcal{X}}}(\mathcal{H}) > M$.

Equivalently, the theorem can be read as a train/test no-free-lunch statement. Fix any hypothesis class $\mathcal{H}$ with $|\mathcal{H}| > 2$, no matter how small, and fix any desired training size M, no matter how large. There exists a partition $\Omega = \mathcal{X}_{\mathrm{tr}} \sqcup \mathcal{X}_{\mathrm{te}}$ with $|\mathcal{X}_{\mathrm{tr}}| = M$ such that no learner succeeds uniformly for every target. This does not contradict the fixed-hierarchy characterizations: countability gives target-wise existence, and finiteness gives target-uniform existence, after $\bar{\mathcal{X}}$ is fixed. The theorem says that changing $\bar{\mathcal{X}}$ can push the required level past any prescribed M.

Complexity-Stratified Bounds. The same question can be asked after stratifying the class by complexity. Once $\bar{\mathcal{X}}$ is fixed, finite sublevel classes give well-defined optimal target-uniform bounds $N^{\bar{\mathcal{X}}}_{(\mathcal{H},c)}(s)$. Can the growth of this function be controlled by $(\mathcal{H}, c)$ alone?

Corollary 11 (Complexity-stratified no free lunch)**.** *Let Ω be countably infinite, let $\mathcal{H} \subseteq \{0,1\}^{\Omega}$ be countably infinite, and let $c : \mathcal{H} \to \mathbb{N}$ have finite sublevel classes. For every nondecreasing $B : \mathbb{N} \to \mathbb{N}_{\geq 1}$ with $B(s) \to \infty$, there exist a hierarchy $\bar{\mathcal{X}}$ on Ω and $s_0 \in \mathbb{N}$ such that, for every $s \geq s_0$,*

$$N^{\bar{\mathcal{X}}}_{(\mathcal{H},c)}(s) > B(s). \tag{13}$$

The key point is quantitative: with $(\mathcal{H}, c)$ fixed, changing the domain hierarchy can make the sublevel domain-complexity function grow faster than any prescribed unbounded rate B. In particular, for some hierarchy, this function has no computable upper bound. In general, hierarchy levels measure task-specific coverage rather than sample size. Both no-free-lunch constructions can reveal one new point per level, so their lower bounds also hold in the number of distinct observed examples.

For length generalization, the hierarchy is the length hierarchy $\bar{\Sigma}$ from Definition 10. A bound for a Transformer-realizable class under this hierarchy is therefore a statement about training on all strings up to a given length and testing on longer strings. It need not transfer to another hierarchy for the same architecture and precision scale. The corollary establishes a worst case over hierarchies, not such a lower bound for any particular natural ordering.

Comparison with Classical NFL. Classical learning-theoretic no-free-lunch theorems are usually read as warnings about hypothesis classes (Shalev-Shwartz & Ben-David, 2014). If the learner is asked to compete with all possible labelings, there is no structure to exploit and no uniform generalization theorem to be had. Our result moves the warning to a second object that is often kept offstage: the structure connecting what is in domain to what is out of domain. Fixing the hypothesis class is therefore not enough. A domain-generalization statement also needs to say what makes the observed part informative about the unobserved part. Counting hypotheses alone is therefore not enough.

7 Conclusion

Our framework connects identification in the limit, length generalization, and online learning while separating their objectives and uniformity requirements. The characterizations distinguish qualitative existence, hierarchy-uniform impossibility, and quantitative bounds for specific classes and hierarchies. This clarifies both the scope of existing guarantees and where further structure is needed.

AI USE STATEMENT

Generative AI tools were used to assist with writing and polishing text, drafting sections of the paper, and developing proofs of mathematical claims. The authors take responsibility for the content of this paper, including its mathematical claims, proofs, and references.

REPRODUCIBILITY STATEMENT

This paper presents theoretical results. The learning framework and assumptions are specified in Section 2, and proofs of the main characterizations, complexity bounds, and impossibility results are provided in Sections B to D. Further supporting arguments and extensions are included in the appendix.

REFERENCES

Emmanuel Abbe, Samy Bengio, Aryo Lotfi, and Kevin Rizk. Generalization on the unseen, logic reasoning and degree curriculum. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31–60. PMLR, 2023. URL https://proceedings.mlr.press/v202/abbe23a.html.

Cem Anil, Yuhuai Wu, Anders Andreassen, Aitor Lewkowycz, Vedant Misra, Vinay Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. Exploring length generalization in large language models. *Advances in Neural Information Processing Systems*, 35:38546–38556, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/fb7451e43f9c1c35b774bcfad7a5714b-Abstract-Conference.html.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems 19*, pp. 137–144. MIT Press, 2006. URL https://papers.nips.cc/paper/2983-analysis-of-representations-for-domain-adaptation.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pp. 41–48. ACM, 2009. doi: 10.1145/1553374.1553380.

Thomas Chen, Tengyu Ma, and Zhiyuan Li. Non-asymptotic length generalization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 8621–8682. PMLR, 2025. URL https://proceedings.mlr.press/v267/chen25ar.html.

E Mark Gold. Language identification in the limit. *Information and Control*, 10(5):447–474, 1967. doi: 10.1016/S0019-9958(67)91165-5.

Xinting Huang, Andy Yang, Satwik Bhattamishra, Yash Sarrof, Andreas Krebs, Hattie Zhou, Preetum Nakkiran, and Michael Hahn. A formal framework for understanding length generalization in transformers. In *International Conference on Learning Representations*, volume 2025, pp. 58095–58179, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/928170bcb050fe64a63fe781b82265aa-Abstract-Conference.html. arXiv:2410.02140.

Zachary Izzo, Eshaan Nichani, and Jason D. Lee. Quantitative bounds for length generalization in transformers. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=TLSUIyBIfs. arXiv:2510.27015.

Jon Kleinberg and Sendhil Mullainathan. Language generation in the limit. In *Advances in Neural Information Processing Systems*, volume 37, pp. 66058–66079. Curran Associates, Inc., 2024. doi: 10.52202/079017-2111. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7988e9b3876ad689e921ce05d711442f-Abstract-Conference.html.

Nayoung Lee, Kartik Sreenivasan, Jason Lee, Kangwook Lee, and Dimitris Papailiopoulos. Teaching arithmetic to small transformers. In *International Conference on Learning Representations*, volume 2024, pp. 25001–25054, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/6bf82fdcbd92b6a7793b3894422d2437-Abstract-Conference.html.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4):285–318, 1988. doi: 10.1023/A:1022869011914.

Vinod Raman, Jiaxun Li, and Ambuj Tewari. Generation through the lens of learning theory. In Nika Haghtalab and Ankur Moitra (eds.), *Proceedings of the Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 4740–4776. PMLR, 2025. URL https://proceedings.mlr.press/v291/raman25a.html.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014. doi: 10.1017/CBO9781107298019.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984. doi: 10.1145/1968.1972.

Vladimir N Vapnik and Alexey Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971. doi: 10.1137/1116025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008. Curran Associates, Inc., 2017. URL https://papers.nips.cc/paper/7181-attention-is-all-you-need.

Andy Yang and David Chiang. Counting like transformers: Compiling temporal counting logic into softmax transformers. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=FmhPg4UJ9K. arXiv:2404.04393.

Andy Yang, Pascal Bergsträßer, Georg Zetzsche, David Chiang, and Anthony W. Lin. Length generalization bounds for transformers. *arXiv preprint arXiv:2603.02238*, 2026a. doi: 10.48550/arXiv.2603.02238. URL https://arxiv.org/abs/2603.02238.

Chenxiao Yang, Nathan Srebro, and Zhiyuan Li. Recursive models for long-horizon reasoning. In *International Conference on Machine Learning*, 2026b. URL https://arxiv.org/abs/2603.02112. arXiv:2603.02112.

Hattie Zhou, Arwen Bradley, Etai Littwin, Noam Razin, Omid Saremi, Joshua Susskind, Samy Bengio, and Preetum Nakkiran. What algorithms can transformers learn? a study in length generalization. In *International Conference on Learning Representations*, volume 2024, pp. 15898–15926, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/45ed1a72597594c097152ef9cc187762-Abstract-Conference.html.

A Auxiliary Definitions

The appendix uses the following derived notation.

For two hypotheses $f, g \in \{0,1\}^\Omega$, write

$$\mathrm{Eq}(f,g) := \{x \in \Omega : f(x) = g(x)\}, \tag{14}$$

also denoted $\{f = g\}$ when this is typographically lighter.

Definition 11 (Version space and pointwise class complexity)**.** *Fix a hypothesis class $\mathcal{H}$, a hierarchy $\bar{\mathcal{X}}$, and a target $f \in \mathcal{H}$. The version space at level n is*

$$V^{\bar{\mathcal{X}}}_{\mathcal{H},n}(f) := \{g \in \mathcal{H} : g|_{\mathcal{X}_n} = f|_{\mathcal{X}_n}\}. \tag{15}$$

The pointwise algorithm-agnostic domain complexity is

$$N^{\bar{\mathcal{X}}}_{\mathcal{H}}(f) := \min\{n \in \mathbb{N}_{\geq 1} : V^{\bar{\mathcal{X}}}_{\mathcal{H},n}(f) = \{f\}\} \tag{16}$$

$$= \min\{n \in \mathbb{N}_{\geq 1} : \forall g \in \mathcal{H} \setminus \{f\},\ \exists x \in \mathcal{X}_n : g(x) \neq f(x)\}. \tag{17}$$

If no such finite level exists, set $N^{\bar{\mathcal{X}}}_{\mathcal{H}}(f) := \infty$.

B Proofs for Exact Characterization

This appendix proves the characterization results from Section 3. The proofs repeatedly use the same elementary observation: if two targets induce the same labeled sample, then any deterministic learner receives the same input under the two targets and must return the same hypothesis.

B.1 Non-Uniform Domain-Dependent Generalization

Proof of Theorem 1. If $\mathcal{H} = \emptyset$, the conclusion is vacuous. Assume henceforth that $\mathcal{H}$ is nonempty. Assume first that $\mathcal{H}$ is countable. Fix an enumeration

$$\mathcal{H} = \{h_1, h_2, \ldots\}. \tag{18}$$

Define the learner A as follows. On any labeled finite sample, if there is at least one hypothesis in $\mathcal{H}$ consistent with that sample, output the least-index consistent hypothesis; if there is none, output h_1. Only the consistent case matters on samples generated by targets in $\mathcal{H}$.

Fix a target $f = h_m$ and a hierarchy $\bar{\mathcal{X}}$. For every $j < m$, the hypotheses h_j and f are distinct, so choose a point $x_j \in \Omega$ such that $h_j(x_j) \neq f(x_j)$. The finite set $\{x_j : j < m\}$ is contained in some prefix $\mathcal{X}_N$ because $\bar{\mathcal{X}}$ exhausts Ω. For every $n \geq N$, each lower-index hypothesis h_j with $j < m$ disagrees with f on $\mathcal{X}_n$, while f itself remains consistent with $f|_{\mathcal{X}_n}$. Hence the least-index consistent hypothesis is exactly f, so $A(f|_{\mathcal{X}_n}) = f$ for all $n \geq N$. Thus $N^{\bar{\mathcal{X}}}_A(f) \leq N$, proving **A1**.

Conversely, suppose $\mathcal{H}$ satisfies **A1**, witnessed by a learner A. Fix a point-by-point hierarchy $\bar{\mathcal{X}}$, say $\mathcal{X}_n = \{x_1, \ldots, x_n\}$. For each $f \in \mathcal{H}$, choose a finite level N_f such that $A(f|_{\mathcal{X}_{N_f}}) = f$, which is possible because $N^{\bar{\mathcal{X}}}_A(f) < \infty$. Consider the map

$$f \longmapsto \big(N_f, f|_{\mathcal{X}_{N_f}}\big). \tag{19}$$

This map is injective. Indeed, if two targets f, g have the same image, then they give A the same labeled sample at the same level; hence

$$f = A(f|_{\mathcal{X}_{N_f}}) = A(g|_{\mathcal{X}_{N_g}}) = g. \tag{20}$$

For each fixed n, there are only finitely many binary labelings of $\mathcal{X}_n$, and there are countably many choices of n. Therefore the range of this injective map is countable, and so $\mathcal{H}$ is countable. □

Proof of Proposition 2. Let $\Omega = \mathbb{N}$ and

$$\mathcal{H} = \{0\} \cup \{h_i : i \geq 1\}, \qquad h_i(j) = \mathbf{1}[i = j], \tag{21}$$

where 0 is the all-zero function. Let the hierarchy be $\mathcal{X}_n = \{1, \ldots, n\}$. Use the least-index consistent learner with 0 listed first.

For the target $f = 0$, every observed prefix is all zero, so 0 is consistent at every level and is listed first. Hence $A(0|_{\mathcal{X}_n}) = 0$ for every $n \geq 1$, and $N_A^{\bar{\mathcal{X}}}(0) = 1$.

However, for every finite n and every $j > n$, the singleton indicator h_j agrees with 0 on $\mathcal{X}_n$. Thus the version space around 0 never collapses at any finite level, even though the learner has already identified the target. □

For the class in Proposition 2, every finite set misses some singleton hypothesis, so the target 0 is not finitely isolated. More generally, finite collapse of the version space is equivalent to the existence of a finite set of domain points that separates the target from all alternatives; this existence does not depend on the ordering of the hierarchy.

Proposition 12 (Finite isolation is hierarchy-invariant)**.** *Fix $f \in \mathcal{H}$. For a hierarchy $\bar{\mathcal{X}}$, write $V_{\mathcal{H},n}^{\bar{\mathcal{X}}}(f) := \{g \in \mathcal{H} : g|_{\mathcal{X}_n} = f|_{\mathcal{X}_n}\}$ for the version space around f at level n. The following are equivalent:*

(i) for some hierarchy $\bar{\mathcal{X}}$, there is a finite level n such that $V_{\mathcal{H},n}^{\bar{\mathcal{X}}}(f) = \{f\}$;
(ii) for every hierarchy $\bar{\mathcal{X}}$, there is a finite level n such that $V_{\mathcal{H},n}^{\bar{\mathcal{X}}}(f) = \{f\}$;
(iii) there is a finite set $S \subseteq \Omega$ such that every $g \in \mathcal{H} \setminus \{f\}$ differs from f on some point of S.

One topological way to understand this is the following. For a finite $S \subseteq \Omega$ and $h \in \mathcal{H}$, let $[h]_S := \{g \in \mathcal{H} : g|_S = h|_S\}$ be the set of hypotheses that agree with h on the finite observation set S. These finite-observation cells generate a topology on $\mathcal{H}$. For any finite-level hierarchy that exhausts Ω, the same topology is generated by the level version spaces $V_{\mathcal{H},n}^{\bar{\mathcal{X}}}(h)$. In this topology, saying that $\{f\}$ is open is exactly saying that some finite observation set isolates f, or equivalently that the version space around f collapses at a finite level.

Proof of Proposition 12. The implication (ii)⇒(i) is immediate. If (i) holds, then any level $\mathcal{X}_n$ with $V_{\mathcal{H},n}^{\bar{\mathcal{X}}}(f) = \{f\}$ is such a finite set S. If (iii) holds, then every hierarchy eventually contains S, and at that level the version space around f is $\{f\}$.

For the topological reformulation, the finite-observation cells $[h]_S$ form a basis: finite intersections of such cells are either empty or another finite-observation cell. Since every finite-level hierarchy that exhausts Ω eventually contains any fixed finite $S \subseteq \Omega$, its level version spaces generate exactly the same topology. Finally, $\{f\}$ is open in this topology if and only if some finite-observation cell $[f]_S$ is equal to $\{f\}$, which is exactly finite isolation. □

The next theorem gives several ways of relaxing the hierarchy quantifier in **A1**: one may fix a hierarchy in advance, or even choose the hierarchy separately for each target. These relaxations have the same threshold because they still ask only for target-by-target success. Thus the following statements are equivalent.

Theorem 13 (Equivalent conditions for **A1**)**.** *The following are equivalent:*

(i) $\exists A\ \forall f \in \mathcal{H}\ \forall \bar{\mathcal{X}}\ \exists N \in \mathbb{N}_{\geq 1} :\ N_A^{\bar{\mathcal{X}}}(f) \leq N$ *(that is, $\mathcal{H}$ satisfies **A1**, Definition 4);*
(ii) $\exists A\ \exists \bar{\mathcal{X}}\ \forall f \in \mathcal{H}\ \exists N \in \mathbb{N}_{\geq 1} :\ N_A^{\bar{\mathcal{X}}}(f) \leq N$*;*
(iii) $\exists A\ \forall f \in \mathcal{H}\ \exists \bar{\mathcal{X}}\ \exists N \in \mathbb{N}_{\geq 1} :\ N_A^{\bar{\mathcal{X}}}(f) \leq N$*;* (iv) *$\mathcal{H}$ is countable.*

Proof of Theorem 13. The implication (i)⇒(ii) follows by fixing any hierarchy, and (ii)⇒(iii) follows by allowing the hierarchy to depend on the target.

Assume (iii). Then there is a learner A such that for every $f \in \mathcal{H}$ there are a hierarchy $\bar{\mathcal{X}}^{\,f}$ and a finite level N_f with

$$A(f|_{\mathcal{X}_{N_f}^{f}}) = f. \tag{22}$$

Assign to f one such finite labeled sample $f|_{\mathcal{X}_{N_f}^{f}}$. This assignment is injective: if two targets are assigned the same labeled finite sample, then A receives the same input and must output both targets.

Since Ω is countable, the collection of finite subsets of Ω is countable, and each finite subset has finitely many binary labelings. Hence there are only countably many finite labeled samples, so $\mathcal{H}$ is countable.

Finally, if $\mathcal{H}$ is countable, then (i) holds by Theorem 1. Therefore all the listed conditions are equivalent. □

B.2 Target-Uniform Domain-Dependent Generalization

Proof of Theorem 3. If $\mathcal{H} = \emptyset$, the conclusion is vacuous. Assume henceforth that $\mathcal{H}$ is nonempty. Assume first that $\mathcal{H}$ is finite. Use the least-index consistent learner from the proof of Theorem 1, for any fixed enumeration of the finite class. Fix a hierarchy $\bar{\mathcal{X}}$. By the argument in Theorem 1, every target $f \in \mathcal{H}$ has a finite identifying level $N_A^{\bar{\mathcal{X}}}(f)$. Since $\mathcal{H}$ is finite,

$$N := \max_{f \in \mathcal{H}} N_A^{\bar{\mathcal{X}}}(f) \tag{23}$$

is finite and works for all targets along this hierarchy. The same learner is used for every hierarchy, so **A2** holds.

Conversely, suppose **A2** holds. Then there are a learner A and, for any fixed hierarchy $\bar{\mathcal{X}}$, a finite level N such that $N_A^{\bar{\mathcal{X}}}(f) \le N$ for every $f \in \mathcal{H}$. If two distinct targets $f, g \in \mathcal{H}$ agreed on $\mathcal{X}_N$, then the learner would receive the same input at level N under both targets. Since $N_A^{\bar{\mathcal{X}}}(f) \le N$ and $N_A^{\bar{\mathcal{X}}}(g) \le N$, it would have to output both f and g on that same input, a contradiction. Thus the restriction map

$$f \longmapsto f|_{\mathcal{X}_N} \tag{24}$$

is injective on $\mathcal{H}$. Because $\mathcal{X}_N$ is finite, it has only finitely many binary labelings. Hence $\mathcal{H}$ is finite. □

Proof of Proposition 4. Fix a hierarchy $\bar{\mathcal{X}}$. If $\mathcal{H} = \emptyset$, the identity is vacuous. Assume henceforth that $\mathcal{H}$ is nonempty.

First suppose a learner A identifies every target in $\mathcal{H}$ by some finite level M, meaning

$$\sup_{f \in \mathcal{H}} N_A^{\bar{\mathcal{X}}}(f) \le M. \tag{25}$$

If two distinct hypotheses $f, g \in \mathcal{H}$ agreed on $\mathcal{X}_M$, then A would receive the same labeled input at level M under both targets. Since both targets must be identified by level M, the same output would have to equal both f and g, impossible. Therefore $\mathcal{X}_M$ distinguishes every pair in $\mathcal{H}$, so

$$N^{\bar{\mathcal{X}}}(\mathcal{H}) \le M. \tag{26}$$

Since this holds for every learner A with a finite uniform identifying level, it gives

$$N^{\bar{\mathcal{X}}}(\mathcal{H}) \le \min_A \sup_{f \in \mathcal{H}} N_A^{\bar{\mathcal{X}}}(f). \tag{27}$$

Conversely, suppose $N^{\bar{\mathcal{X}}}(\mathcal{H}) = N < \infty$. Define a learner that, on a labeled finite sample, outputs the unique hypothesis in $\mathcal{H}$ consistent with that sample if such a unique hypothesis exists, and outputs an arbitrary fixed element of $\mathcal{H}$ otherwise. For every target $f \in \mathcal{H}$ and every $n \ge N$, the prefix $\mathcal{X}_n$ contains $\mathcal{X}_N$, and $\mathcal{X}_N$ distinguishes every pair in $\mathcal{H}$. Hence f is the unique hypothesis in $\mathcal{H}$ consistent with $f|_{\mathcal{X}_n}$, so the learner outputs f. Therefore

$$\min_A \sup_{f \in \mathcal{H}} N_A^{\bar{\mathcal{X}}}(f) \le N^{\bar{\mathcal{X}}}(\mathcal{H}). \tag{28}$$

If $N^{\bar{\mathcal{X}}}(\mathcal{H}) = \infty$, the first paragraph shows that no learner can have a finite uniform identifying level. Thus both sides are ∞. This proves the identity in all cases. □

Proof of Theorem 5. The implication (i)⇒(ii) is immediate by fixing one hierarchy. If (ii) holds with learner A, hierarchy $\bar{\mathcal{X}}$, and bound N, then no two distinct targets can agree on $\mathcal{X}_N$: otherwise

A would receive the same labeled input under both targets at level N, but would have to output two different functions. Hence $\mathcal{X}_N$ distinguishes every pair in $\mathcal{H}$, so $N^{\bar{\mathcal{X}}}(\mathcal{H}) < \infty$, proving (iii).

If (iii) holds, then for some hierarchy $\bar{\mathcal{X}}$ and some finite N, the prefix $\mathcal{X}_N$ distinguishes every pair in $\mathcal{H}$. The restriction map $f \mapsto f|_{\mathcal{X}_N}$ is injective, so $\mathcal{H}$ injects into the finite set of binary labelings of $\mathcal{X}_N$. Hence $\mathcal{H}$ is finite, proving (v).

If $\mathcal{H}$ is finite, then for every hierarchy $\bar{\mathcal{X}}$ and every distinct pair $f, g \in \mathcal{H}$, there is some point $x_{f,g}$ with $f(x_{f,g}) \neq g(x_{f,g})$. Since $\bar{\mathcal{X}}$ exhausts Ω, each such point appears in some finite prefix. Taking the maximum over the finitely many pairs gives a finite level whose prefix distinguishes every pair in $\mathcal{H}$. Thus (iv) holds. If $\mathcal{H}$ has at most one element, the distinguishing condition is vacuous and any level, say 1, works.

Finally, (iv)⇒(iii) is immediate, and (v)⇒(i) also follows from Theorem 3. Therefore all five conditions are equivalent. □

B.3 Domain-Uniform Generalization

Proof of Theorem 6. The implication (i)⇒(ii) is immediate, since a single bound that works for all targets and all hierarchies is in particular allowed to depend on the target.

Assume next that Ω is finite. Under the finite-domain convention, every hierarchy has revealed all of Ω by level $|\Omega|$. Consider the learner that outputs the unique hypothesis in $\mathcal{H}$ consistent with the full observed labels once the full domain has been revealed, and is arbitrary before that. For any target $f \in \mathcal{H}$, the full-domain labeled sample determines f uniquely inside $\mathcal{H}$. Hence this learner identifies every target along every hierarchy by level $|\Omega|$, proving **A4**, and therefore **A3**.

It remains to prove that **A3** implies Ω is finite, under the assumption $|\mathcal{H}| > 2$. Suppose **A3** is witnessed by a learner A. For every $f \in \mathcal{H}$, choose a finite level bound N_f such that

$$N_A^{\bar{\mathcal{X}}}(f) \leq N_f \tag{29}$$

for every hierarchy $\bar{\mathcal{X}}$.

We first show that every pair of distinct hypotheses has a finite agreement set. Let $f, g \in \mathcal{H}$ be distinct and set $M = \max\{N_f, N_g\}$. If $\mathrm{Eq}(f, g)$ contained at least M points, choose a hierarchy whose M-th prefix consists of M such agreement points. At level M, the labeled samples generated by f and by g are identical. But $M \geq N_f, N_g$, so domain-uniform identifiability requires A to output f on this input under target f and g on the same input under target g, impossible. Therefore $\mathrm{Eq}(f, g)$ has fewer than M points, and is finite.

Now choose three distinct hypotheses $h_1, h_2, h_3 \in \mathcal{H}$. Since the labels are binary, at each point $x \in \Omega$ at least two of the three values $h_1(x), h_2(x), h_3(x)$ are equal. Hence

$$\Omega = \mathrm{Eq}(h_1, h_2) \cup \mathrm{Eq}(h_1, h_3) \cup \mathrm{Eq}(h_2, h_3). \tag{30}$$

Each set on the right is finite by the previous paragraph, so Ω is finite. This proves (ii)⇒(iii), and completes the equivalence. □

C Proofs for Stratified Domain Complexity

Proof of Theorem 7. Fix a hierarchy $\bar{\mathcal{X}}$, a class $\mathcal{H} \subseteq \{0,1\}^\Omega$, and a complexity scale $c : \mathcal{H} \to \mathbb{N}$.

Fix $s \in \mathbb{N}$. By Equation (6), $\mathcal{H}^c_{\leq s}$ is finite. For every distinct pair $f, g \in \mathcal{H}^c_{\leq s}$, choose a point $x_{f,g}$ with $f(x_{f,g}) \neq g(x_{f,g})$. Since $\bar{\mathcal{X}}$ exhausts Ω, each such point appears in some finite prefix. Taking the maximum over the finitely many pairs gives a finite prefix that distinguishes every pair in $\mathcal{H}^c_{\leq s}$. If $|\mathcal{H}^c_{\leq s}| \leq 1$, any level, say 1, works. Thus

$$N^{\bar{\mathcal{X}}}_{(\mathcal{H},c)}(s) < \infty. \tag{31}$$

For the optimality identity, fix s with $\mathcal{H}^c_{\leq s} \neq \emptyset$. By definition,

$$N^{\bar{\mathcal{X}}}_{(\mathcal{H},c)}(s) = N^{\bar{\mathcal{X}}}(\mathcal{H}^c_{\leq s}). \tag{32}$$

Applying Proposition 4 to the class $\mathcal{H}^c_{\leq s}$ gives

$$N^{\bar{\mathcal{X}}}_{(\mathcal{H},c)}(s) = \min_A \sup_{f \in \mathcal{H}^c_{\leq s}} N^{\bar{\mathcal{X}}}_A(f), \tag{33}$$

with both sides taking values in $\mathbb{N}_{\geq 1} \cup \{\infty\}$.

Finally suppose $|\mathcal{H}^c_{\leq s}| \to \infty$. The sublevel classes are nested in s, so the numbers $N^{\bar{\mathcal{X}}}_{(\mathcal{H},c)}(s)$ are nondecreasing. If they did not tend to infinity, then there would be a finite M and arbitrarily large s with $N^{\bar{\mathcal{X}}}_{(\mathcal{H},c)}(s) \leq M$. For each such s, the prefix $\mathcal{X}_M$ would distinguish every pair in $\mathcal{H}^c_{\leq s}$, so the restriction map into $\{0,1\}^{\mathcal{X}_M}$ would be injective on $\mathcal{H}^c_{\leq s}$. This gives

$$|\mathcal{H}^c_{\leq s}| \leq 2^{|\mathcal{X}_M|} \tag{34}$$

for arbitrarily large s, contradicting $|\mathcal{H}^c_{\leq s}| \to \infty$. Therefore $N^{\bar{\mathcal{X}}}_{(\mathcal{H},c)}(s) \to \infty$. □

Proof of Proposition 8. Suppose, toward a contradiction, that for some computable exhaustive hierarchy $\bar{\mathcal{X}}$ there is a computable function B satisfying the bound in the statement. Given two descriptions $p, q \in \Pi$, compute

$$s := \max\{\kappa(p), \kappa(q)\}. \tag{35}$$

Compute $B(s)$ and enumerate $\mathcal{X}_{B(s)}$. Since evaluation is decidable from descriptions, we can check whether h_p and h_q disagree on some point of $\mathcal{X}_{B(s)}$. If they do, then they are not semantically equivalent. If they do not, then h_p and h_q agree on $\mathcal{X}_{B(s)}$. But $h_p, h_q \in \mathcal{H}^{\Pi,\kappa}_{\leq s}$, and $B(s)$ is at least the sublevel domain complexity of this slice. Hence any two distinct hypotheses in $\mathcal{H}^{\Pi,\kappa}_{\leq s}$ would have been separated inside $\mathcal{X}_{B(s)}$. Therefore $h_p = h_q$ on all of Ω, and the procedure decides semantic equivalence from descriptions. □

Proof of Proposition 9. Fix an enumeration $z_1, z_2, \ldots$ of Ω. We construct the hierarchy recursively. Set $\mathcal{X}_0 = \emptyset$. Given $\mathcal{X}_{m-1}$, for each $i < m$ choose a point $w_{i,m} \in \Omega$ such that $h_i(w_{i,m}) \neq h_m(w_{i,m})$. Let u_m be the first point in the enumeration of Ω that is not contained in

$$\mathcal{X}_{m-1} \cup \{w_{i,m} : i < m\}. \tag{36}$$

Define

$$\mathcal{X}_m := \mathcal{X}_{m-1} \cup \{w_{i,m} : i < m\} \cup \{u_m\}. \tag{37}$$

Each $\mathcal{X}_m$ is finite and $\mathcal{X}_{m-1} \subsetneq \mathcal{X}_m$. The added points u_m ensure that every point of Ω eventually appears, so $\bar{\mathcal{X}}$ is exhaustive. Finally, for any pair $i < j \leq m$, the point $w_{i,j}$ was added by level j, and hence belongs to $\mathcal{X}_m$; it separates h_i from h_j. Thus $\mathcal{X}_m$ distinguishes every pair in $\{h_1, \ldots, h_m\}$, so $N^{\bar{\mathcal{X}}}(\{h_1, \ldots, h_m\}) \leq m$. □

D Proofs for the No-Free-Lunch Results

D.1 Class-Level No Free Lunch

Proof of Theorem 10. Choose three distinct hypotheses $h_1, h_2, h_3 \in \mathcal{H}$. Since labels are binary, the three pairwise agreement sets

$$\mathrm{Eq}(h_1, h_2), \qquad \mathrm{Eq}(h_1, h_3), \qquad \mathrm{Eq}(h_2, h_3) \tag{38}$$

cover Ω: at every point, at least two of the three binary labels agree. Since Ω is infinite, at least one of these three agreement sets is infinite. Let $g_0, g_1 \in \mathcal{H}$ be a distinct pair with $\mathrm{Eq}(g_0, g_1)$ infinite.

Fix $M \in \mathbb{N}_{\geq 1}$. Choose distinct points

$$z_1, \ldots, z_M \in \mathrm{Eq}(g_0, g_1), \tag{39}$$

and extend them to an enumeration $z_1, z_2, \ldots$ of Ω. Define the point-by-point hierarchy

$$\mathcal{X}_n := \{z_1, \ldots, z_n\}. \tag{40}$$

Then g_0 and g_1 agree on $\mathcal{X}_M$.

Now let A be any learner. At level M, the labeled sample generated by target g_0 is the same as the labeled sample generated by target g_1. Therefore A has the same output on both samples. It cannot be equal to both distinct functions g_0 and g_1. Hence at least one of the two targets, call it h, is not identified by level M, so

$$N_A^{\bar{\mathcal{X}}}(h) > M. \tag{41}$$

The hierarchy was constructed before choosing A, so this proves the first claim.

When $\mathcal{H}$ is finite, the equivalent class-domain-complexity statement follows from Proposition 4. Indeed,

$$N^{\bar{\mathcal{X}}}(\mathcal{H}) > M \tag{42}$$

if and only if no learner can identify every target in $\mathcal{H}$ by level M, which is equivalent to saying that every learner has some target $h \in \mathcal{H}$ with $N_A^{\bar{\mathcal{X}}}(h) > M$. □

D.2 Complexity-Stratified No Free Lunch

Proof of Corollary 11. Fix an enumeration

$$\Omega = \{y_1, y_2, \ldots\}. \tag{43}$$

We will construct a point-by-point hierarchy by constructing an enumeration $z_1, z_2, \ldots$ of Ω. The construction ensures that, for infinitely many increasing complexity thresholds, a pair of hypotheses inside the corresponding sublevel class agrees on a very long initial prefix.

Step 1: a reusable extension claim. We first state the combinatorial step used at every stage. Let $P \subseteq \Omega$ be finite, let $S \in \mathbb{N}$, and let $R \subseteq \Omega$ be infinite. Since $\mathcal{H}^c_{\leq S}$ is finite and $\mathcal{H}$ is infinite, the set $\mathcal{H} \setminus \mathcal{H}^c_{\leq S}$ is infinite. Only finitely many binary label patterns can occur on P, so infinitely many hypotheses in $\mathcal{H} \setminus \mathcal{H}^c_{\leq S}$ share the same pattern on P. Choose three such hypotheses. On every point of R, at least two of their three binary labels agree; since the three pairwise agreement sets cover R, one pair agrees on infinitely many points of R. Thus there exist distinct $f, g \in \mathcal{H} \setminus \mathcal{H}^c_{\leq S}$ such that

$$f|_P = g|_P \qquad \text{and} \qquad \mathrm{Eq}(f, g) \cap R \text{ is infinite.} \tag{44}$$

Step 2: build a hierarchy with longer and longer concealed pairs. Choose any three hypotheses in $\mathcal{H}$. Their three pairwise agreement sets cover the infinite set Ω, so one pair (f_0, g_0) has infinite agreement set. Let

$$A_0 := \mathrm{Eq}(f_0, g_0), \qquad S_0 := \max\{c(f_0), c(g_0)\}, \qquad P_0 := \emptyset. \tag{45}$$

Inductively, suppose we have a finite prefix P_k of the ordering, a pair (f_k, g_k), an infinite agreement set $A_k := \mathrm{Eq}(f_k, g_k)$, and a threshold $S_k := \max\{c(f_k), c(g_k)\}$, with $P_k \subseteq A_k$. The invariant means that the current pair agrees on every point revealed so far.

Apply the combinatorial step with the finite set $P_k \cup \{y_{k+1}\}$, the threshold S_k, and the infinite set A_k. We obtain a pair (f_{k+1}, g_{k+1}) outside $\mathcal{H}^c_{\leq S_k}$ such that it agrees on $P_k \cup \{y_{k+1}\}$ and has infinitely many agreement points inside A_k. Set

$$A_{k+1} := \mathrm{Eq}(f_{k+1}, g_{k+1}), \qquad S_{k+1} := \max\{c(f_{k+1}), c(g_{k+1})\}. \tag{46}$$

Because both f_{k+1} and g_{k+1} lie outside $\mathcal{H}^c_{\leq S_k}$, we have $S_{k+1} > S_k$. Also $A_k \cap A_{k+1}$ is infinite, and $P_k \cup \{y_{k+1}\} \subseteq A_{k+1}$.

Extend the current ordering by adding unused points from $A_k \cap A_{k+1}$ until at least the first $B(S_{k+1})$ positions have been filled. Then append y_{k+1} if it has not already appeared. Let P_{k+1} be the resulting finite prefix. This preserves the induction invariant $P_{k+1} \subseteq A_{k+1}$: the old prefix and y_{k+1} lie in A_{k+1}, and the added points lie in $A_k \cap A_{k+1}$. It also ensures that the first $B(S_{k+1})$ points of the ordering lie in A_k.

Since y_{k+1} is inserted by the end of stage $k+1$ if it has not appeared earlier, every point of Ω eventually appears. The limiting ordering $z_1, z_2, \ldots$ is therefore an enumeration of Ω, and it defines a point-by-point hierarchy $\bar{\mathcal{X}}$ by

$$\mathcal{X}_n := \{z_1, \ldots, z_n\}. \tag{47}$$

Step 3: verify the lower bound on the sublevel domain complexity. It remains to verify the bound. The thresholds S_k are strictly increasing, hence $S_k \to \infty$. Let $s \geq S_0$. Choose k such that

$$S_k \leq s \leq S_{k+1}. \tag{48}$$

The pair (f_k, g_k) lies in $\mathcal{H}^c_{\leq s}$ because both complexities are at most $S_k \leq s$. By construction, f_k and g_k agree on the first $B(S_{k+1})$ points of the hierarchy. Since B is nondecreasing and $s \leq S_{k+1}$, they agree on the first $B(s)$ points. Therefore $\mathcal{X}_{B(s)}$ does not distinguish all pairs in $\mathcal{H}^c_{\leq s}$, and by the definition of sublevel domain complexity,

$$N^{\bar{\mathcal{X}}}_{(\mathcal{H},c)}(s) > B(s). \tag{49}$$

Thus the corollary holds with $s_0 = S_0$. □

E Extended Discussions

We consider four extensions: online mistake bounds, approximate prediction, distributional support mismatch, and positive-only presentations.

E.1 Online Learning

Online learning keeps the same hierarchy, but replaces the identifying level by a mistake count. Before level n is revealed, the prediction rule has observed the labels on $\mathcal{X}_{n-1}$ and must predict the labels on the new increment $\Delta_n = \mathcal{X}_n \setminus \mathcal{X}_{n-1}$. We count one mistake at level n if any of these predictions is wrong.

This is weaker than identification. It counts errors only when they are exposed on a newly revealed increment; identification requires eventual stabilization on a predictor correct on all of Ω.

As in Definition 2, the online prediction rule may output an arbitrary binary predictor in $\{0,1\}^\Omega$, not necessarily an element of $\mathcal{H}$.

Definition 12 (Uniform mistake bound)**.** *For an online prediction rule A, target $f \in \mathcal{H}$, and hierarchy $\bar{\mathcal{X}}$, define the* mistake count

$$M_A(f, \bar{\mathcal{X}}) := \sum_{n \geq 1} \mathbf{1}\left[A(f|_{\mathcal{X}_{n-1}})|_{\mathcal{X}_n \setminus \mathcal{X}_{n-1}} \not\equiv f|_{\mathcal{X}_n \setminus \mathcal{X}_{n-1}}\right], \tag{50}$$

where $\mathcal{X}_0 = \emptyset$. We say $\mathcal{H}$ admits a uniform mistake bound *if there exist an online prediction rule A and a finite $M \in \mathbb{N}$ such that $M_A(f, \bar{\mathcal{X}}) \leq M$ for every $f \in \mathcal{H}$ and every hierarchy $\bar{\mathcal{X}}$.*

Let $\mathrm{Ldim}(\mathcal{H})$ denote the Littlestone dimension of $\mathcal{H}$, namely the supremum of the depths of complete binary trees shattered by $\mathcal{H}$. The classical mistake-bound theorem of Littlestone (1988) applies to the criterion above.

Theorem 14 (Littlestone's mistake-bound theorem)**.** *$\mathcal{H}$ admits a uniform mistake bound if and only if $\mathrm{Ldim}(\mathcal{H}) < \infty$. Moreover, the optimal mistake bound is exactly $\mathrm{Ldim}(\mathcal{H})$.*

Thus Littlestone dimension controls the number of mistaken revealed increments. It does not control the level at which the target has been identified.

From Mistakes to Identification. The mistake-bound theorem is uniform over both targets and hierarchies, just as **A4** is. The difference is what the bound promises: **A4** gives a level after which the target has been identified, while a mistake bound only limits how many revealed increments can be predicted incorrectly. The next proposition places this online criterion between the finite and countable cases characterized in Section 3.

Proposition 15 (Intermediate mistake complexity)**.** *Assume Ω is countable. Then*

$$|\mathcal{H}| < \infty \quad \Longrightarrow \quad \mathrm{Ldim}(\mathcal{H}) < \infty \quad \Longrightarrow \quad \mathcal{H} \text{ is countable}. \tag{51}$$

Thus, over a countable domain, finite Littlestone dimension sits between finiteness and countability. The same counting argument extends to any infinite domain: if $\mathrm{Ldim}(\mathcal{H}) = d < \infty$, then $|\mathcal{H}| \leq |\Omega|$.

Fix a well-order of Ω and run a bounded-mistake learner; each target is determined by its at-most-d mistake locations. Since an infinite Ω has only $|\Omega|$ subsets of size at most d, the bound follows.

Combining Proposition 15 with Theorems 1, 3 and 14 gives

$$\textbf{A2} \quad \Longrightarrow \quad \text{uniform mistake bound} \quad \Longrightarrow \quad \textbf{A1}. \tag{52}$$

Thus online learning is weaker than target-uniform identification, even when the same mistake bound must work for every hierarchy. A mistake bound controls how many times disagreement can be revealed as error, but not how long the hierarchy can postpone the first point that distinguishes the remaining possible targets.

Example 1 (One mistake does not imply target-uniform identification)**.** *Let $\Omega = \{x_1, x_2, \ldots\}$ and consider the class of singleton indicators*

$$\mathcal{H}_{\text{sing}} := \{h_i : i \geq 1\}, \qquad h_i(x_j) = \mathbf{1}[i = j]. \tag{53}$$

*Then $\mathrm{Ldim}(\mathcal{H}_{\text{sing}}) = 1$, but $\mathcal{H}_{\text{sing}}$ does not satisfy **A2**.*

For this class, online prediction is easy: predict zero until the unique positive point appears. At most one mistake is possible. Identification is different. Along the natural hierarchy $\mathcal{X}_N = \{x_1, \ldots, x_N\}$, all targets h_i with $i > N$ have the same all-zero labels at level N. Thus no common finite level can force the learner to know which singleton target is correct.

E.2 Approximation

A second comparison keeps the hierarchy but weakens exact identification to small P-error. So far, success meant exact identification: after some level, the learner must recover the target on all of Ω. Here we only ask for small error under a fixed evaluation distribution. The question is whether the hierarchy-uniform impossibility from exact identification persists under this weaker requirement.

Fix a distribution P on Ω. For a target $f \in \mathcal{H}$ and predictor $h \in \{0,1\}^\Omega$, the P-error of h relative to f is

$$L_{P,f}(h) := P(\{x \in \Omega : h(x) \neq f(x)\}). \tag{54}$$

Definition 13 (ε-approximate domain complexity)**.** *For $\varepsilon > 0$, learner A, hierarchy $\bar{\mathcal{X}}$, evaluation distribution P, and target $f \in \mathcal{H}$, define*

$$N_{A,\varepsilon}^{\bar{\mathcal{X}},P}(f) := \min\{i \in \mathbb{N}_{\geq 1} : \forall n \geq i,\ L_{P,f}(A(f|_{\mathcal{X}_n})) \leq \varepsilon\}, \tag{55}$$

with value ∞ if no such finite i exists. This is the first level after which every later predictor has P-error at most ε.

For a fixed hierarchy, approximation has a simple tail-mass guarantee: every consistent learner satisfies

$$L_{P,f}(A(f|_{\mathcal{X}_n})) \leq P(\Omega \setminus \mathcal{X}_n). \tag{56}$$

Because the hierarchy is exhaustive, $P(\Omega \setminus \mathcal{X}_n) \to 0$. Thus fixed-hierarchy approximation does not require any combinatorial structure of $\mathcal{H}$. The nontrivial question is whether one can choose a level bound before the hierarchy is known. Say that P has full support if $P(x) > 0$ for every $x \in \Omega$. Then every disagreement between distinct hypotheses has positive P-mass: if $u \neq v$, then $P(\{u \neq v\}) > 0$.

Proposition 16 (Approximate no free lunch over hierarchies)**.** *Let Ω be countably infinite, let P have full support on Ω, and let $\mathcal{H} \subseteq \{0,1\}^\Omega$ have $|\mathcal{H}| > 2$. Then there exists $\varepsilon > 0$ such that, for every $M \in \mathbb{N}_{\geq 1}$, there is a hierarchy $\bar{\mathcal{X}}$ for which every learner A has some target $f \in \mathcal{H}$ satisfying*

$$N_{A,\varepsilon}^{\bar{\mathcal{X}},P}(f) > M. \tag{57}$$

Thus allowing error does not produce a hierarchy-independent rate. For a fixed hierarchy and fixed evaluation distribution, consistency gives convergence in P-error. But before the hierarchy is fixed, the points on which hypotheses differ can be postponed for arbitrarily many levels, even when those points have positive P-mass.

E.3 Out-of-Support Generalization

A third variant returns to a distributional observation model. The learner is no longer given a hierarchy of observed regions; it receives i.i.d. labeled samples from a training distribution Q, returns a binary predictor, and is evaluated under a possibly different distribution P. The hierarchy-uniform impossibility results above do not apply literally to this model. The same issue reappears as a support question: does training ever reveal the distinctions that evaluation can test?

The support of Q is the region from which labels can ever be observed. Sampling reveals points in $\mathrm{supp}(Q)$ randomly, with repetitions, but never reveals labels outside that support. We measure error by the same P-error as in Section E.2.

Full-Support Generalization. If Q has full support, meaning $Q(x) > 0$ for every $x \in \Omega$, no point is permanently outside training. Every point that can affect the P-error is at least possible to observe. For finite classes, this gives a target-uniform sample size.

Theorem 17 (Full-support generalization)**.** *Let Ω be countably infinite, let Q have full support on Ω, fix any evaluation distribution P on Ω, and let $\mathcal{H} \subseteq \{0,1\}^{\Omega}$ be finite. Then there exists a learner A such that, for every $\varepsilon, \delta > 0$, there is a finite N such that, for every target $f \in \mathcal{H}$ and every $n \geq N$,*

$$\mathbb{P}_{(x_1,\ldots,x_n)\sim Q^n}\left[L_{P,f}\big(A\big((x_i, f(x_i))_{i=1}^n\big)\big) \leq \varepsilon\right] \geq 1 - \delta. \tag{58}$$

Full-support sampling removes the permanent hiding place for any fixed disagreement set. Since a finite class has only finitely many pairwise disagreement sets, one sample size can make all of them appear with high probability.

Out-of-Support. If Q is not full support, some labels are never observed. The next theorem says that this gap can break the same target-uniform conclusion even when Q has infinite support: the learner may see infinitely many possible training points, but still not the part of Ω on which evaluation distinguishes two targets.

Theorem 18 (Out-of-support generalization can fail)**.** *Let Ω be countably infinite, let P have full support on Ω, and let $\mathcal{H} \subseteq \{0,1\}^{\Omega}$ with $|\mathcal{H}| \geq 3$. Then there exist a distribution Q with infinite support and an $\varepsilon > 0$ such that, for every learner A, every $N \in \mathbb{N}_{\geq 1}$, and every $n \geq N$, there exists a target $f \in \mathcal{H}$ satisfying*

$$\mathbb{P}_{(x_1,\ldots,x_n)\sim Q^n}\left[L_{P,f}\big(A\big((x_i, f(x_i))_{i=1}^n\big)\big) > \varepsilon\right] \geq \frac{1}{2}. \tag{59}$$

Thus the relevant condition is not how many points Q can sample from. It is whether the training support includes the distinctions that P-error will test. Full support guarantees this; an infinite support that is not full need not.

E.4 Positive-Only Presentations

Finally, consider a nearby observation model, familiar from identification from text and generation in the limit (Gold, 1967; Kleinberg & Mullainathan, 2024), in which the learner sees only positive examples. For $f \in \mathcal{H}$, write

$$\mathrm{Pos}(f) := \{x \in \Omega : f(x) = 1\} \tag{60}$$

for its positive region. A positive presentation of f is a strictly increasing sequence $\emptyset = S_0 \subsetneq S_1 \subsetneq S_2 \subsetneq \cdots$ of finite subsets of $\mathrm{Pos}(f)$ whose union is $\mathrm{Pos}(f)$. We restrict to the nontrivial case in which every $\mathrm{Pos}(f)$ is infinite. Write $\mathrm{Pres}^+(f)$ for the set of all such presentations.

We ask for the same fully uniform exact-identification guarantee as before, but with positive sets replacing labeled finite domains. Here A denotes a positive-only learner whose input is a finite positive set:

$$\exists A\ \exists N < \infty\ \forall f \in \mathcal{H}\ \forall \bar{S} \in \mathrm{Pres}^+(f) :\ \forall t \geq N,\ A(S_t) = f. \tag{61}$$

The relevant condition is finite pairwise positive overlap: distinct targets cannot share arbitrarily many positive points.

Proposition 19 (Uniform positive-only identification)**.** *Assume* $\mathrm{Pos}(f)$ *is infinite for every* $f \in \mathcal{H}$*. Then the fully uniform positive-only criterion in Equation* (61) *holds if and only if*

$$\sup_{\substack{f,g\in\mathcal{H}\\ f\neq g}} |\mathrm{Pos}(f) \cap \mathrm{Pos}(g)| < \infty, \tag{62}$$

with the convention that the supremum over the empty set is 0*. Moreover, if this supremum is* d*, then* $N = d + 1$ *suffices.*

The condition is not finiteness. For example, take $\Omega = \mathbb{N} \times \mathbb{N}$ and, for $a, b \in \mathbb{N}$, let

$$G_{a,b} := \{(t, at + b) : t \in \mathbb{N}\}, \tag{63}$$

and let $f_{a,b}$ be the indicator of $G_{a,b}$. The class $\{f_{a,b} : a, b \in \mathbb{N}\}$ is countably infinite, and every positive region is infinite. Yet two distinct graphs intersect in at most one point, so two positive observations already certify the target against every competitor.

In our terms, this restricts the presentation to positive examples while keeping exact identification and full uniformity. Under this restriction, the strongest uniform statement can hold even for infinite classes.

This positive-only model is closely related to generation in the limit (Kleinberg & Mullainathan, 2024). There too, the learner sees a growing finite subset of the target language, but only needs to generate one new valid positive example. Since the target is unknown, that point must be positive for every consistent hypothesis; closure dimension characterizes when this can be done uniformly (Raman et al., 2025). The proposition above shows that the stronger goal of recovering the whole language, equivalently generating all positives, can also hold uniformly.

F Proofs for Online Learning

Proof of Theorem 14. We first prove the upper bound when $d := \mathrm{Ldim}(\mathcal{H}) < \infty$, using the convention $\mathrm{Ldim}(\emptyset) = -1$. The learner maintains the version space $V_n \subseteq \mathcal{H}$ of hypotheses consistent with all labels revealed before level n. Thus $V_1 = \mathcal{H}$, and after level n is revealed the version space is restricted to hypotheses consistent with the new labels.

For a point x and a label $b \in \{0, 1\}$, write

$$V_n(x, b) := \{h \in V_n : h(x) = b\}. \tag{64}$$

Before seeing the labels on $\mathcal{X}_n \setminus \mathcal{X}_{n-1}$, the learner predicts at each x a label b maximizing $\mathrm{Ldim}(V_n(x, b))$, breaking ties arbitrarily. After all labels at level n are revealed, it updates the version space to the subset of V_n consistent with the whole increment.

We claim that every mistaken level strictly decreases the Littlestone dimension of the version space. Suppose level n is mistaken. Then for some $x \in \mathcal{X}_n \setminus \mathcal{X}_{n-1}$, the learner predicted $1 - b$ while the true label is $b = f(x)$. Since the learner chose a label of maximum Littlestone dimension,

$$\mathrm{Ldim}(V_n(x, 1 - b)) \geq \mathrm{Ldim}(V_n(x, b)). \tag{65}$$

Both $V_n(x, 0)$ and $V_n(x, 1)$ are subclasses of V_n, so their Littlestone dimensions are at most $\mathrm{Ldim}(V_n)$. If $\mathrm{Ldim}(V_n(x, b)) = \mathrm{Ldim}(V_n)$, then also $\mathrm{Ldim}(V_n(x, 1 - b)) = \mathrm{Ldim}(V_n)$, and placing x at the root above two shattered trees of depth $\mathrm{Ldim}(V_n)$ would shatter a tree of depth $\mathrm{Ldim}(V_n) + 1$, a contradiction. Hence

$$\mathrm{Ldim}(V_n(x, b)) < \mathrm{Ldim}(V_n). \tag{66}$$

The updated version space after the whole increment is contained in $V_n(x, b)$, so its Littlestone dimension is also strictly smaller than $\mathrm{Ldim}(V_n)$. Since the dimension starts at d and never goes below zero, at most d levels can be mistaken.

For the lower bound, let d be any integer such that $\mathcal{H}$ shatters a complete binary tree of depth d. We show that every deterministic learner can be forced to make d mistakes; if $\mathrm{Ldim}(\mathcal{H}) = \infty$, this holds for every finite d.

The adversary follows a shattered tree. At the root, it presents the root instance as the first singleton increment. If the learner predicts 0, the adversary reveals label 1; if the learner predicts 1, it reveals

label 0. The adversary then moves to the child corresponding to the revealed label and repeats this procedure for d levels. By shattering, the resulting root-to-leaf sequence of labels is realized by some target $f \in \mathcal{H}$.

The points along the realized path are distinct. If the same point appeared twice on one root-to-leaf path, then at the later occurrence the two outgoing branches would require both labels while the earlier occurrence has already fixed one label, contradicting shattering. Thus the path points can be used as distinct singleton increments of a hierarchy, which is then completed arbitrarily to exhaust Ω. By construction, the learner makes one mistake at each of the first d levels. Therefore no learner can have a uniform mistake bound smaller than d. Taking the supremum over all shattered depths gives the lower bound $\mathrm{Ldim}(\mathcal{H})$. □

Proof of Proposition 15. If $\mathcal{H}$ shatters a complete binary tree of depth d, then the 2^d root-to-leaf label sequences must be realized by 2^d distinct hypotheses: two paths that first diverge at a node assign different labels to the same instance at that node. Hence, when $\mathcal{H}$ is finite,

$$2^d \le |\mathcal{H}| \tag{67}$$

for every shattered depth d, so $\mathrm{Ldim}(\mathcal{H}) \le \log_2 |\mathcal{H}| < \infty$.

Now suppose $\mathrm{Ldim}(\mathcal{H}) = d < \infty$. By Theorem 14, there is a deterministic online prediction rule that makes at most d mistakes on every target and hierarchy. Fix a singleton hierarchy enumerating Ω, say $\mathcal{X}_n = \{x_1, \dots, x_n\}$.

For each target $f \in \mathcal{H}$, let $E_f \subseteq \mathbb{N}$ be the finite set of times at which this learner makes a mistake when run against f on the fixed hierarchy. Then $|E_f| \le d$. The set E_f determines f: given E_f, replay the learner from the beginning; at time n, compute its prediction from the labels reconstructed so far, and set the true label to be the prediction if $n \notin E_f$ and the opposite label if $n \in E_f$. This reconstructs $f(x_n)$ for every n. Since the hierarchy enumerates Ω, it reconstructs f on all of Ω.

Thus the map $f \mapsto E_f$ is injective from $\mathcal{H}$ into the collection of finite subsets of $\mathbb{N}$ of size at most d, which is countable. Therefore $\mathcal{H}$ is countable. □

Verification of Example 1. The class shatters a one-point tree: at any point x_j, the hypothesis h_j realizes label 1, while any h_i with $i \neq j$ realizes label 0. Hence $\mathrm{Ldim}(\mathcal{H}_{\mathrm{sing}}) \ge 1$.

It cannot shatter a depth-two tree. If the root is labeled by some point x_j and the root label is 1, then the only consistent target is h_j. All later labels are then fixed, so the next node cannot realize both outgoing labels. Thus no depth-two tree is shattered, and $\mathrm{Ldim}(\mathcal{H}_{\mathrm{sing}}) = 1$.

Equivalently, the all-zero online predictor makes at most one mistake on this class, namely when the unique positive point of the target singleton is revealed.

Finally, consider the singleton hierarchy $\mathcal{X}_n = \{x_1, \dots, x_n\}$. Suppose **A2** held along this hierarchy with some learner and common level N. For every $i > N$, the target h_i gives the same all-zero labeled sample on $\mathcal{X}_N$. At level N, the learner would therefore receive the same input for all targets h_i with $i > N$, but would have to output each of these distinct targets. This is impossible. Hence no target-uniform identification bound exists, so **A2** fails. □

G Proofs for Approximation

Proof of Proposition 16. Choose three distinct hypotheses in $\mathcal{H}$. Since the labels are binary, the three pairwise agreement sets cover Ω; because Ω is infinite, some pair f, g agrees on infinitely many points. Since P has full support and $f \neq g$,

$$\Delta := P(\{f \neq g\}) > 0. \tag{68}$$

Set $\varepsilon := \Delta/3$. Given M, choose a finite set $S \subseteq \{f = g\}$ of size M, list these points first, and complete the list to a hierarchy with $\mathcal{X}_M = S$. Since $f|_S = g|_S$, the learner returns the same predictor on the two training sets; call it $\hat{h}$. On every point where f and g disagree, $\hat{h}$ is wrong for at least one of the two targets. Hence

$$\Delta \le L_{P,f}(\hat{h}) + L_{P,g}(\hat{h}). \tag{69}$$

Therefore at least one of the two targets, say $u \in \{f, g\}$, has $L_{P,u}(A(u|_{\mathcal{X}_M})) \geq \Delta/2 > \varepsilon$, and so $N_{A,\varepsilon}^{\tilde{\mathcal{X}},P}(u) > M$. □

H Proofs for Out-of-Support Generalization

Proof of Theorem 17. Enumerate the finite class as $\mathcal{H} = \{h_1, \ldots, h_M\}$ and use the least-index consistent learner: given a finite labeled sample, output the first hypothesis in the enumeration consistent with all observed labels, if one exists, and output h_1 otherwise. The learner ignores the order and multiplicity of repeated sample points.

If $M = 1$, the claim is immediate. Assume $M \geq 2$. For every pair of distinct hypotheses $h_j, h_k \in \mathcal{H}$, define

$$D_{j,k} := \{x \in \Omega : h_j(x) \neq h_k(x)\}. \tag{70}$$

Each $D_{j,k}$ is nonempty and has positive Q-mass because Q has full support. Let

$$\alpha := \min_{j \neq k} Q(D_{j,k}) > 0. \tag{71}$$

If the sample contains at least one point from every pairwise disagreement set, then the observed labels distinguish the target from every other hypothesis in $\mathcal{H}$. The least-index consistent learner therefore outputs the target, and the P-error is zero.

By a union bound, the probability of missing some pairwise disagreement set is at most

$$M(M-1)(1-\alpha)^n, \tag{72}$$

which tends to zero as $n \to \infty$. Hence for every $\delta > 0$ there is a finite N such that, for all $n \geq N$, the sample hits every pairwise disagreement set with probability at least $1 - \delta$. On this event, the learner outputs the target for every $f \in \mathcal{H}$, and therefore has P-error at most ε for every target. This gives the target-uniform bound. □

Proof of Theorem 18. Choose three distinct hypotheses in $\mathcal{H}$. Their three pairwise agreement sets cover the infinite domain Ω, so some pair $u, v \in \mathcal{H}$ has an infinite agreement set $\mathrm{Eq}(u, v)$. Choose any distribution Q with infinite support contained in $\mathrm{Eq}(u, v)$; for example, enumerate countably many distinct points in $\mathrm{Eq}(u, v)$ and assign them probabilities proportional to 2^{-k}.

Since $u \neq v$, the disagreement set $\{u \neq v\}$ is nonempty. Because P has full support,

$$\Delta := P(\{u \neq v\}) > 0. \tag{73}$$

Choose $\varepsilon < \Delta/2$.

Fix a learner A and a sample size n. Samples from Q always lie in $\mathrm{Eq}(u, v)$, so the labeled sample given to A is identical under target u and target v. For any fixed sample sequence, let $\hat{h}$ be this common learner output. At each point where u and v disagree, $\hat{h}$ is wrong for at least one of the two targets. Therefore

$$L_{P,u}(\hat{h}) + L_{P,v}(\hat{h}) \geq \Delta. \tag{74}$$

Thus, for every sample sequence, at least one of the two targets has error larger than ε. Averaging over Q^n, at least one of the two targets has failure probability at least $1/2$. Since A and n were arbitrary, this holds for every learner and every sample size; in particular, it holds for every N and every $n \geq N$. □

I Proofs for Positive-Only Presentations

Proof of Proposition 19. First suppose

$$d := \sup_{\substack{f,g \in \mathcal{H} \\ f \neq g}} |\mathrm{Pos}(f) \cap \mathrm{Pos}(g)| < \infty. \tag{75}$$

Define a learner as follows. Given a finite positive set S, if there is a unique $h \in \mathcal{H}$ such that $S \subseteq \mathrm{Pos}(h)$, output that h; otherwise output an arbitrary binary predictor.

Fix a target $f \in \mathcal{H}$ and a positive presentation $\bar{S} \in \mathrm{Pres}^{+}(f)$. Since the presentation is increasing and every step adds a positive point, $|S_t| \geq t$. Hence $|S_t| \geq d+1$ for every $t \geq d+1$. No $g \neq f$ can contain all points in S_t, because then $|\mathrm{Pos}(f) \cap \mathrm{Pos}(g)| \geq |S_t| \geq d+1$, contradicting the definition of d. Thus f is the unique hypothesis whose positive region contains S_t, and the learner outputs f for all $t \geq d+1$.

Conversely, suppose

$$\sup_{\substack{f,g \in \mathcal{H} \\ f \neq g}} |\mathrm{Pos}(f) \cap \mathrm{Pos}(g)| = \infty. \tag{76}$$

Let A be any learner and let $N < \infty$. Choose $f \neq g$ with $|\mathrm{Pos}(f) \cap \mathrm{Pos}(g)| \geq N$, and choose distinct points $x_1, \ldots, x_N \in \mathrm{Pos}(f) \cap \mathrm{Pos}(g)$. Since both positive regions are infinite, we can extend the common initial sequence

$$S_t = \{x_1, \ldots, x_t\}, \qquad 1 \leq t \leq N, \tag{77}$$

to positive presentations of both f and g. At time N, the learner receives the same input under the two targets. It therefore gives the same output in both worlds, and cannot be correct for both f and g. No learner can satisfy Equation (61) with this N, and since N was arbitrary, no uniform finite bound exists. □

An infinite example. The finite-overlap condition can hold for infinite classes. Let $\Omega = \mathbb{N} \times \mathbb{N}$ and, for $a, b \in \mathbb{N}$, define

$$G_{a,b} := \{(t, at + b) : t \in \mathbb{N}\}, \tag{78}$$

and let $f_{a,b}$ be the indicator of $G_{a,b}$. Each $\mathrm{Pos}(f_{a,b}) = G_{a,b}$ is infinite, and the class $\{f_{a,b} : a, b \in \mathbb{N}\}$ is countably infinite. Two distinct such sets intersect in at most one point, since two distinct affine functions over $\mathbb{N}$ agree at at most one input. Hence the pairwise positive overlap is at most one, and two positive observations uniformly identify the target.

J Limitations and Scope

Our results are information-theoretic: they characterize when identification is possible, not how to compute the identifying level for a given $(\mathcal{H}, c, \bar{\mathcal{X}})$. Three structural restrictions are worth noting. First, we work in the realizable, deterministic-labeling setting; agnostic and noisy presentations are deferred to future work. Second, we focus on exact identification; approximate analogues with ε-error are discussed in Section E.2. Third, the framework is set-indexed rather than sequence-indexed, hiding within-level order; online-learning variants are discussed in Section E.1. The work is theoretical and does not raise direct societal concerns.